Towards a genomic approach to plant authentication and quality control

Félix Forest, William Baker, Steven Dodsworth, Olivier Maurin, Lisa Pokorny, Laura Botigue, Jan Kim, Niroshini Epitalawage, Grace Brewer, Vanessa Barber, and Monique S. Simmonds
Kew’s Plant and Fungal Trees of Life

Aims to “generate and compile high-throughput sequencing data for one representative of all 14,000 flowering plant genera and all ca 8,000 fungal genera” by 2020.
Sampling all angiosperm genera

- Living Collections
- DNA Bank
- Seed bank
- Tissue Bank

43% missing
57% available

ca 95% of angiosperms genera in Kew’s herbarium

Funding:
- The Calleva Foundation
- The Sackler Trust
- The Garfield Weston Foundation
What about fungi?

- Still largely unknown! **1.5-3 (10?) million fungal species.**
- The 1000 Fungal Genome Project only targeted family level, and focused mainly on fungi that are easy to grow in pure culture.
- **Whole Genome Sequencing:** Average genome size in fungi ~50MB (ranging 10MB-800MB) compared to ~3GB in plants.
- Bait set currently in development at Kew.

---

**8194** accepted fungal genera in Index Fungorum.

**6000** of the accepted genera (many >30 years old) are found in Kew’s Fungarium

---

**Tales from the crypt: genome mining from fungarium specimens improves resolution of the mushroom tree of life**

BRYN T. M. DENTINGER1,2†, ESTER GAYA1†, HEATH O’BRIEN3, LAURA M. SUZ1, ROBERT LACHLAN4, JORGE R. DÍAZ-VALDERRAMA5, RACHEL A. KOCH6 and M. CATHERINE AIME5

1Jodrell Laboratory, Royal Botanic Gardens, Kew, TW9 3DS, UK
Next-generation DNA barcoding

Potential HTS methods for DNA authentication (not an exhaustive list!):

- **Amplicon sequencing**: numerous barcodes sequenced at smaller cost than using Sanger, but gain in discrimination could be limited and has limitations similar to Sanger sequencing for degraded material (e.g. amplification of degraded DNA).
- **Transcriptomics**: costly and would require fresh material from all parts of the plant.
- **Whole genome sequencing**: possible for some groups, like fungi, but not for plants in which genome size varies 2400-fold.
- **Plastid genome sequencing**: Given the use of plastid sequences as standard barcodes, this is an obvious choice in order to integrate existing data.
- **Targeted enrichment**: would allow the selection of a set of nuclear regions (several 100s) and potentially greatly improve discriminatory power.
- **Genome skimming**: would recover plastid genome and ribosomal, thus all regions previously used. Relatively low cost if highly multiplexed, but would result in low coverage and patchy results that may be difficult to compare between samples and with reference data.
The Hyb-Seq approach, “the combination of target enrichment and genome skimming, allows simultaneous data collection for low-copy nuclear genes and high-copy genomic targets for plant systematics and evolution studies”

Using RNA baits to selectively enrich target loci

A.K.A. – Bait hybridisation, target enrichment, sequence capture, anchored phylogenomics, Hyb-Seq...

Most commonly manufactured by MYcroarray – MYbaits – custom and off-the-shelf kits available for different applications
The ideal angiosperm-wide baits

- Manageable set of nuclear loci that can be confidently align across deep divergences

- One simple kit that works equally well across angiosperms – *breadth*

- Utility at multiple scales (i.e. also useful for species-level phylogenetics, which is the most common scale of work) – *depth*

- Effective with suboptimal template DNA (i.e. degraded DNA from herbarium specimens, processed material)
Angiosperm-wide baits

Starting data:
• 1KP transcriptome data (209 angiosperm families, ~600 genera)
• Alignments for 410 single or low-copy nuclear genes in all plants

Finding representative sequences:
• Bait hybridisation tolerates 15-25% divergence
• Select a number of representative sequences per gene in order to stay within 25% divergence threshold.

Results:
• 170 genes with no divergent sequence
• 118 genes with <1% divergent sequence
• ca 50 genes were removed as > 25% divergence (these genes can not be reliably recovered with baits at this scale)

• Allowing up to 5% divergent sequences => a final set of 353 genes (75,151 baits)
• Total alignment length = 295,713 bp
Retrieval of loci across 192 samples

- Samples spread across all orders of angiosperms
- Good recovery of targets across the set
- Some genes appear to be underperforming
- Some samples did not perform well
- Effect of sequencing depth on number of loci recovered to be explored
• Phylogenetic/taxonomic representation of loci retrieval
• No obvious biases
• Perhaps Rosids performing slightly better but could be sample size artefact
• Distance to number of bait sequences yet to be analysed
Completing the pilot study

Initial pilot study of the baits
• Almost completed, one last 96-well plate to be sequenced.

Completing the APG IV families
• 1KP = 209 families
• Pilot: +207 families
• Aim: all 416 families completed; about 20 short at the moment.

Priority samples
• 25% of genera in each families
• Focus on some particular families such as Solanaceae, Iridaceae, Asteraceae, Fabaceae, Poaceae, Orchidaceae, Sapindaceae, Arecaceae.
Species-level variation

*Nicotiana sect Suaveolentes (Solanaceae)*
- Comprises about 40 very closely related species.
- Plastid sequence data has been so far unhelpful in resolving species relationships
- Low sequence divergence
- Angiosperm-wide baits has worked well with the seven samples included in the pilot study.

- *Babiana* (Iridaceae) – 8 accessions from a small clade of closely related species
- *Magnolia* (Magnoliaceae) – 8 species representing different sections/groups in the genus
- *Nymphaea* (Nymphaeaceae) – the entire genus
- *Nepenthes* (Nepenthaceae) – 7 species representing different sections/groups in the genus.
- *Basselinia* (Arecales) – rapid radiation of 14 species, endemic to New Caledonia.
The ideal approach?

- **Hyb-Seq combines genome skimming and targeted enrichment**, thus sequencing the traditional barcodes as well as several 100s more from both the plastid and nuclear genomes. Also use of repeat sequences.

- **Standard across angiosperms**, i.e. a bait set that works for all angiosperms.

- **Suboptimal DNA quality** obtained from herbarium specimens can be used for producing reference data, and degraded and processed material from the trade would be usable.

- **Relatively affordable**:
  - Bait kit provided by Arbor Bioscience (former Mycroarray) as off-the-shelf kit for US$ 125 per reaction.
  - At Kew, £50 per sample; one company US£ 165 per sample.

- **Captures all plant barcodes used to date**, i.e. the official *rbcL* and *matK*, and others that have been used such as *psbA-trnH*, as well as ITS (often ITS2 only) and the P6 loop of *trnL*.

- Thus, the Hyb-Seq approach advocated here will also allow the combination of existing reference datasets produced by studies using different barcodes.

- Adopting this approach doesn’t mean the end of standard barcodes as both approaches are compatible and support the proposal of Coissac et al 2016 (i.e. “twin track approach”; Mol Ecol 25:1423).
Phylogenomics of Detarioideae (Leguminosae)
Manuel de la Estrella, RBG Kew.

- 350 samples, 250 from herbarium specimens of varying ages and preservation quality.
- ca 95% success; others with reduced set of loci, but still useful.
The ideal approach?

- **Hyb-Seq combines genome skimming and targeted enrichment**, thus sequencing the traditional barcodes as well as several 100s more from both the plastid and nuclear genomes. Also use of repeat sequences.

- **Standard across plants**, i.e. a bait set that works for all plants.

- **Suboptimal DNA quality** obtained from herbarium specimens can be used for producing reference data, and degraded and processed material from the trade would be usable.

- **Relatively affordable**:
  - Bait kit provided by Arbor Bioscience (former Mycroarray) as off-the-shelf kit for US$ 125 per reaction.
  - At Kew, £50-60 per sample; one company US£ 165 per sample.

- **Captures all plant barcodes used to date**, i.e the official *rbcL* and *matK*, and others that have been used such as *psbA-trnH*, as well as ITS (often ITS2 only) and the P6 loop of *trnL*.

- Thus, the targeted enrichment approach advocated here will also **allow the combination of existing reference datasets** produced by studies using different barcodes.

- Adopting this approach doesn’t mean the end of standard barcodes, as both **approaches are compatible**, and it support the proposal of Coissac et al 2016 (i.e. ”twin track approach”; Mol Ecol 25:1423).
Some overlap with existing bait sets.
As for standard barcodes, this will facilitate the integration of existing reference data.
Any problems?

• Some of the problems of standard barcodes won’t go away with HTS barcoding such as hybridization.
• Cost, in the short term at least.
• Data storage requirements significantly larger than standard barcodes.
• Effect of genome size?
• Accessibility and ease of analysis.
• Expandable to all land plants?
Potential genome size effect

- Extremely large genomes do seem to be underperforming (50 pg upwards)
- All giant genomes had much lower recovery of targets
- Much variability in recovery at within the “standard” range of genome sizes < 5 pg
Any problems?

• Some of the problems of standard barcodes won’t go away with HTS barcoding such as hybridization.
• Cost, in the short term at least.
• Data storage requirements significantly larger than standard barcodes.
• Effect of genome size?
• Accessibility and ease of analysis.
• Expandable to all land plants?
A bioinformatic challenge

• Pipeline for regularly building trees using newly generated genomic data and available data in GenBank, including existing barcodes.
• Trees entirely or partially downloadable, depending on user needs
• Interactive trees for different audiences, annotated with data suitable for different audiences – the PAFTOL Explorer
• Addition of a specific barcoding module?
Any problems?

• Some of the problems of standard barcodes won’t go away with HTS barcoding such as hybridization.
• Cost, in the short term at least.
• Data storage requirements significantly larger than standard barcodes.
• Effect of genome size?
• Accessibility and ease of analysis.
• Expandable to all land plants?
Earth BioGenome Project

A massive global effort is proposed, the Earth BioGenome Project (EBP), to sequence and characterize the genomes of all known eukaryotic species on earth in ten years.
Acknowledgements

The angiosperm-wide bait set should be made public by the end of 2017 !!

• 1KP project (onekp.com).
• Pam & Doug Soltis, Jim Leebens-Mack.
• Matt Johnson and Norm Wickett (bait design)
• Many colleagues at RBG Kew and Kew Foundation

Funding:
• The Calleva Foundation
• The Sackler Trust
• The Garfield Weston Foundation